CLAIMS

1. A computer assisted method of auditing a superset of training data, the superset comprising examples of documents having one or more category assignments, the method including:

partitioning the superset into at least two disjoint sets, including a test set and a training set, wherein the test set includes one or more test documents and the training set includes examples of documents belonging belong to at least two categories;

categorizing the test documents using the training set;

calculating a metric of confidence based on results of the categorizing step and the category assignments for the test documents; and

reporting the test documents and category assignments that are suspicious and that appear to be missing, based on the metric of confidence.

2. The method of claim 1, further including repeating the partitioning, categorizing and calculating steps until at least one-half of the documents in the superset have been assigned to the test set.

3. The method of claim 2, wherein the test set created in the partition step has a single test document.

4. The method of claim 2, wherein the test set created in the partition step has a plurality of test documents.

5. The method of claim 1, further including repeating the partitioning, categorizing and calculating steps until substantially all of the documents in the superset have been assigned to the test set.

6. The method of claim 1, wherein the partitioning, categorizing and calculating steps are carried out substantially without user intervention.

7. The method of claim 5, wherein the partitioning, categorizing and calculating steps are carried out substantially without user intervention.

8. The method of claim 1, wherein the partitioning, categorizing, calculating and reporting steps are carried out substantially without user intervention.

9. The method of claim 5, wherein the partitioning, categorizing, calculating and reporting steps are carried out substantially without user intervention.

10. The method of claim 1, wherein the categorizing step includes determining k nearest neighbors of the test documents and the calculating step is based on a k nearest neighbors categorization logic.

11. The method of claim 10, wherein the metric of confidence is an unweighted measure of distance between the test document and the examples of documents belonging to various categories.

12. The method of claim 11, where the unweighted measure includes application of a relationship $\Omega_0(\mathbf{d}_t, T_m) = \sum_{\mathbf{d} \in \{K(\mathbf{d}_t) \cap T_m\}} s(\mathbf{d}_t, \mathbf{d})$, wherein

Ω_0 is a function of the test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and

s is a metric of distance between the test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}$, the certain sample documents being among a set of k nearest neighbors of the test document having category assignments to the various categories T_m.

13. The method of claim 10, wherein the metric of confidence is a weighted measure of distance between the test document and the examples of documents belonging to various categories, the weighted measure taking into account the density of a neighborhood of the test document.

14. The method of claim 13 where the weighted measure includes application of a relationship $\Omega_1(\mathbf{d}_t, T_m) = \frac{\sum_{\mathbf{d}_1 \in \{K(\mathbf{d}_t) \cap T_m\}} s(\mathbf{d}_t, \mathbf{d}_1)}{\sum_{\mathbf{d}_2 \in K(\mathbf{d}_t)} s(\mathbf{d}_t, \mathbf{d}_2)}$, wherein

Ω_1 is a function of the test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and

s is a metric of distance between the test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, the certain sample documents $\mathbf{d}_1$ being among a set of k nearest neighbors of the test document having category assignments to the various categories T_m and the certain sample documents $\mathbf{d}_2$ being among a set of k nearest neighbors of the test document.

15. The method of claim 1, wherein the identifying step further includes filtering the test documents based on the metric of confidence.

16. The method of claim 15, wherein the filtering step further includes color coding the identified test documents based on the metric of confidence.

17. The method of claim 15, wherein the filtering step further includes selecting for display the identified test documents based on the metric of confidence.

18. The method of claim 1, wherein the user interface is a printed report.

19. The method of claim 1, wherein the user interface is a file conforming to XML syntax.

20. The method of claim 1, wherein the user interface is a sorted display identifying at least a portion of the test documents.

21. The method of claim 1, further including calculating a precision score for the identified test documents.

22. A computer assisted method of auditing a superset of training data, the superset comprising examples of documents having one or more category assignments, the method including:

determining k nearest neighbors of the documents in the superset;

categorizing the documents based on the k nearest neighbors into a plurality of categories;

calculating a metric of confidence based on results of the categorizing step and the category assignments for the documents; and

reporting the documents and category assignments that are suspicious and that appear to be missing, based on the metric of confidence.

23. The method of claim 22, wherein the metric of confidence is an unweighted measure of distance between the test document and the examples of documents belonging to various categories.

24. The method of claim 23, where the unweighted measure includes application of a relationship $\Omega_0(\mathbf{d}_t, T_m) = \sum_{\mathbf{d}\in\{K(\mathbf{d}_t)\cap T_m\}} s(\mathbf{d}_t, \mathbf{d})$, wherein

Ω_0 is a function of the test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and

s is a metric of distance between the test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}$, the certain sample documents being among a set of k nearest neighbors of the test document having category assignments to the various categories T_m.

25. The method of claim 22, wherein the metric of confidence is a weighted measure of distance between the test document and the examples of documents belonging to various categories, the weighted measure taking into account the density of a neighborhood of the test document.

26. The method of claim 25, wherein the weighted measure includes application of a relationship $\Omega_1(\mathbf{d}_t, T_m) = \frac{\sum_{\mathbf{d}_1\in\{K(\mathbf{d}_t)\cap T_m\}} s(\mathbf{d}_t, \mathbf{d}_1)}{\sum_{\mathbf{d}_2\in K(\mathbf{d}_t)} s(\mathbf{d}_t, \mathbf{d}_2)}$, wherein

Ω_1 is a function of the test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and

s is a metric of distance between the test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, the certain sample documents $\mathbf{d}_1$ being among a set of k nearest neighbors of the test document having category assignments to the various categories T_m and the certain sample documents $\mathbf{d}_2$ being among a set of k nearest neighbors of the test document.

27. The method of claim 22, wherein the determining, categorizing and calculating steps are carried out substantially without user intervention.

28. The method of claim 22, wherein the identifying step further includes filtering the documents based on the metric of confidence.

29. The method of claim 28, wherein the filtering step further includes color coding the identified documents based on the metric of confidence.

30. The method of claim 28, wherein the filtering step further includes selecting for display the identified documents based on the metric of confidence.

31. The method of claim 22, wherein the user interface is a printed report.

32. The method of claim 22, wherein the user interface is a file conforming to XML syntax.

33. The method of claim 22, wherein the user interface is a sorted display identifying at least a portion of the documents.

34. The method of claim 22, further including calculating a precision score for the identified documents.